Assignment 2: Transformer Architecture

Introduction

In this assignment, I compare three transformer architectures for text summarization: **GPT-2** (decoder-only), **BERT** (encoder-only), and **T5** (encoder-decoder). The goal is to analyze how architectural differences affect generative performance, training behavior, and summary quality. I use the CNN/DailyMail dataset to train and test all three models for the same summarization task.

Methodology

1. Dataset

I used the CNN/DailyMail dataset (v3.0.0) because it is a common standard benchmark for text summarization. The task involves creating concise summaries from long news articles, making it appropriate for assessing generative transformer models.

Dataset Size: 500 training samples, 100 validation samples.

Preprocessing:

- Articles were tokenized using specific model tokenizers.
- Maximum input length: **512** tokens.
- Maximum summary length: **128** tokens.
- For GPT-2 and T5, text was formatted for conditional summarization.
- For BERT, articles were split into sentences for extractive classification.

2. Models

Model	Architecture	Approach
GPT-2	Decoder-only	**Causal language modeling**: It generates summaries word-by-word based on a prompt, like auto-complete.
BERT	Encoder-only	**Sentence classification**: It selects the most important sentences from the article without adding any new text.
T5	Encoder-decoder	**Text-to-text transfer**: It reads the entire article, then generates a completely new summary, like a translation.

3. Training Strategies

Model	Strategy	Batch Size	Epochs	Learning Rate	Objective
GPT-2	Prompt-based generation and mask prompts in loss	2	1	5e-5	Autoregressive causal language modeling (predict next token).
BERT	Sentence-level classification using ROUGE-based labels	16	1	2e-5	Extractive binary classification.
T5	Text-to-text conditional generation with summarization prefix.	8	2	5e-5	Sequence-to-sequence generation (encoder–decoder).

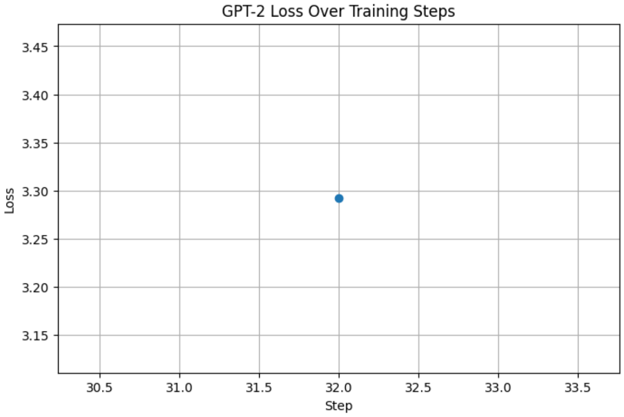
Results

1. Quantitative Results

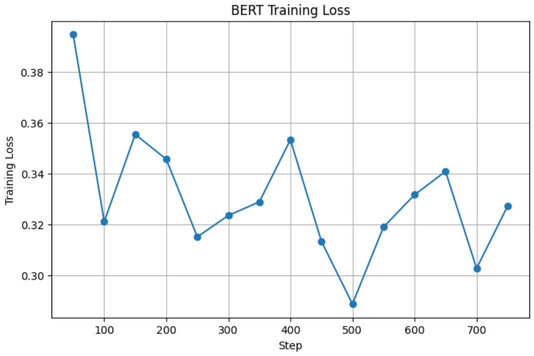
Metric	GPT-2	BERT	T5
Training Loss	3.92	0.302	2.436
Eval Loss	2.944	0.341	2.053
Perplexity	18.2	Nil	7.79
ROUGE-1	0.228	0.364	0.403
ROUGE-2	0.115	0.146	0.182
ROUGE-L	0.166	0.227	0.275

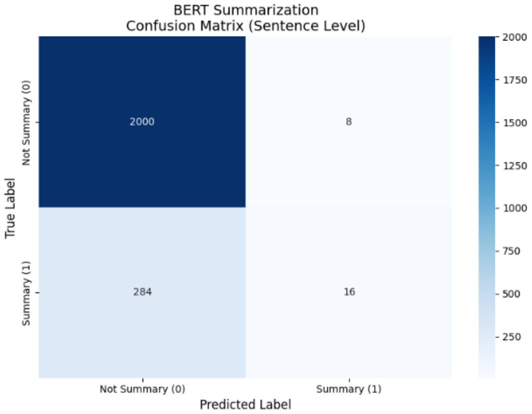
- BERT’s low loss is expected because it performs a simpler task. It only needs to decide whether a sentence is important or not (binary classification). GPT-2 and T5 face the more difficult challenge of producing novel text. T5's lower eval loss indicates better generalization than GPT-2.
- T5 also has much lower perplexity (7.79 compared to 18.2 for GPT-2). Lower perplexity means the model is more confident and better at predicting the next word. This happens because T5 first reads and understands the entire article (using its encoder) and then generates the summary.
- T5 achieves the highest score across all metrics despite having the fewest parameters. This confirms that encoder-decoder architectures are optimal for conditional generation tasks. BERT's competitive ROUGE-1 (0.364) shows extractive methods can be effective, but its lower ROUGE-2 indicates limited ability to capture novel phrases.

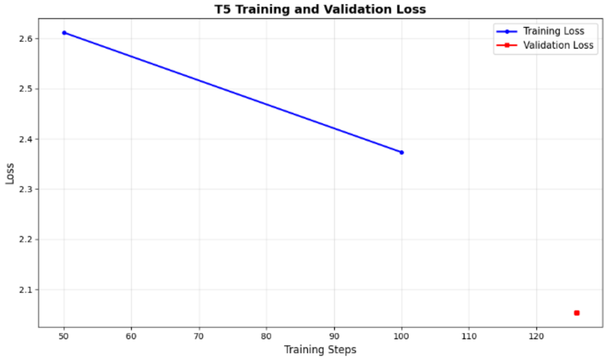
2. Visualization



3.







Qualitative Results

Example 1

Reference Summary: Jarryd Hayne quit the NRL in October to try and get into American Football .This week, he signed a three-year contract with the San Francisco 49ers . The chairman of the US Association of Rugby League welcomed his arrival .

GPT-2: Jarryd Hayne's move to the NFL is a boost for rugby league in the United States, it has been claimed. The Australian full-back or centre quit the National Rugby League in October to try his luck in American football and was this week given a three-year contract with the 49ers. Peter Illfield, chairman of US Association of Rugby League, said: 'Jarryd

BERT: Jarryd Hayne's move to the NFL is a boost for rugby league in the United States, it has been claimed. The Australia international full-back or centre quit the National Rugby League in October to try his luck in American football and was this week given a three-year contract with the San Francisco 49ers. Jarryd Hayne (right) has signed with the San Francisco 49ers after quitting the NRL in October .

T5: Jarryd Hayne has signed a three-year contract with the San Francisco 49ers. the Australian full-back or centre quit the National Rugby League in October to try his luck in American football and is expected to play at full back and centre in rugby league, according to the chairman of US Association of Rugby league.

Analysis

The training and comparison of the three transformer architectures revealed significant differences in stability, performance, and suitability for summarization tasks.

GPT-2: The decoder-only model was simple to implement but challenging to train effectively. It generates text in an autoregressive manner and does not explicitly model full conversational context. Despite producing fluent text, it had a high training loss (3.9) and the lowest ROUGE scores of the three models. While GPT-2 can perform summarization, it is not specifically designed for this purpose. It needs a lot more training to get good at structured tasks like summarization compared to models built specifically for that purpose.

BERT: It had the lowest training loss (0.30) and concluded quickly due to its simple binary classification objective. However, its encoder-only architecture limits it to extractive summarization. It selects key sentences from the article rather than creating new text. While this approach maintains data accuracy and performs well on ROUGE-1, it does not have the ability to paraphrase, compress, or organize information across sentences.

T5: It performed best despite having fewer parameters. Its encoder-decoder architecture is optimized for text-to-text tasks, allowing for efficient generation. The sequential encoder extracts global context from the entire article, while the autoregressive decoder produces accurate summaries based on that representation. As a result, T5 had the highest ROUGE scores and the lowest perplexity (7.79), which means improved generalization and more confident predictions.

Chain-of-Thought (CoT) Reflection

Chain-of-thought (CoT) prompting encourages models to generate intermediate reasoning steps before producing a final output. Although this assignment focused on summarization rather than explicit reasoning tasks, architectural differences suggest varying suitability for CoT. **For GPT-2 (decoder-only),** CoT prompting could help structure outputs, but because it is autoregressive, it may also increase repetition or complexity. Without bidirectional context, reasoning chains may become unstable. **For BERT (encoder-only),** it does not benefit from CoT because the model is not generative. BERT is good at contextual encoding and classification, but it cannot generate step-by-step reasoning text. **For T5 (encoder-decoder),** CoT is likely to be the most effective. Its encoder captures all contextual information, while its decoder produces structured outputs. As a result, encoder-decoder architectures are better suited for tasks that require multi-step reasoning or structured generation. Overall, CoT is most naturally aligned with encoder-decoder models that are intended for conditional text generation.

Conclusion

The results show that architectural design has a significant impact on summarization task performance. Encoder-decoder models, such as T5, are ideal for conditional generation because they explicitly separate context encoding and output generation. Encoder-only models, such as BERT, excel at understanding and classification but have limited generative capabilities. Decoder-only models, such as GPT-2, excel at open-ended text generation but struggle with structured summarization. Overall, the encoder-decoder architecture performed best for this task, which shows that model structure is more important than parameter count when solving conditional generation problems.